Number agreement in copular constructions. A treebank-based investigation

May 15, 2014

This paper has both a theoretical and a methodological objective. The theoretical one concerns the modeling of number agreement in copular constructions. For that purpose it adopts the distinction, familiar from Head-driven Phrase Structure Grammar, between morpho-syntactic agreement (aka concord) and index agreement. The methodological objective concerns the demonstration of how treebanks can be exploited in order to guide the formulation of relevant generalizations. For that purpose we crucially rely on tools and resources that have recently been developed in the framework of the Dutch-Flemish STEVIN programme (2004-2011) and the European CLARIN programme (2009-2015).

1 Introduction

Copular constructions minimally consist of a subject, a copular verb and a subject-oriented predicative complement. In such constructions there is not only number agreement between the verb and the subject, but also between the predicative complement and the subject, as illustrated for Italian in (1).[1]

(1) a. Questo cane è molto forte.
this dog.SG is.SG very strong.SG
'This dog is very strong.'

b. Questi cani sono molto forti.
these dog.PL are.PL very strong.PL
'These dogs are very strong.'

[1]There is also person agreement between the subject and the verb and gender agreement between the subject and the predicate adjective, but this paper focuses on number agreement.

Mismatches, however, are not excluded, as demonstrated by the following French examples, quoted from Wechsler and Zlatić (2003, 98,102).

(2) a. Vous êtes loyal.
you.PL be.PL loyal.SG
'You are loyal.'

b. On a été loyaux.
one.SG have.SG been loyal.PL
'We have been loyal.'

While the verbs show number agreement with the subject, the predicative complements do not: If the plural *vous* is understood to denote an individual, rather than an aggregate, the predicative complement is singular, as in (2a), and if the singular *on* is understood to denote and aggregate, rather than an individual, the predicative complement is plural, as in (2b).

The challenge for a treatment of these data is to make it sufficiently restrictive to enforce agreement wherever it is required and sufficiently flexible to allow mismatches as those in (2). A treatment that aims to meet this challenge is the one of Head-driven Phrase Structure Grammar (HPSG). It is based on a distinction between two types of agreement.

1.1 Concord vs. index agreement

The distinction between morpho-syntactic agreement (also known as concord) and index agreement was introduced in Pollard and Sag (1994) (chapter 2) and further developed in a.o. Kathol (1999) and Wechsler and Zlatić (2003).[2] The latter defines it in terms of the scheme in (3).

(3) morphology $\Longleftrightarrow$ CONCORD $\Longleftrightarrow$ INDEX $\Longleftrightarrow$ semantics

"We recognize two distinct grammaticalization 'portals', one each via semantics and morphology. These two sources of grammaticalization lead to two distinct bundles of agreement features for a given noun. The morphology-related agreement bundle will be called CONCORD (which includes case, number and gender) and the semantics-related agreement bundle which will be called INDEX (which includes person, number and gender)." (Wechsler and Zlatić 2003, 28)

[2] A similar distinction has been proposed in transformational grammar. Sauerland and Elbourne (2002), for instance, employs the NUMBER feature to model morpho-syntactic agreement, and a new feature, called MEREOLOGY, to capture something which closely resembles index agreement.

For most nouns, the number and gender features in the two 'portals' match, but if there is a mismatch between morphology and semantics, as in the case of a morpho-syntactically plural pronoun with a single referent, the NUMBER value in the index may reflect the latter and deviate from the former. This is made explicit in the lexical entry that Kathol (1999, 248) assigns to *vous*.[3]

(4)
$$\begin{bmatrix} \ldots \mid \text{AGR} \begin{bmatrix} \text{NUMBER} & \textit{plural} \\ \text{GENDER} & \textit{gender} \end{bmatrix} \\ \ldots \mid \text{INDEX} \begin{bmatrix} \text{PERSON} & 2 \\ \text{NUMBER} & \textit{number} \\ \text{GENDER} & \textit{gender} \end{bmatrix} \end{bmatrix}$$

The AGR|NUMBER value is unambiguously *plural*, but its counterpart in the index is left underspecified. This accounts for (2a), if one assumes that the agreement between subject and verb is an instance of concord, while the agreement between subject and predicative complement is an instance of index agreement, as spelled out in (5), quoted from Kathol (1999, 241).[4]

(5) a. morpho-syntactic: AGR(selector) $\approx$ AGR(argument)
b. semantic: AGR(selector) $\approx$ INDEX(argument)

"$\approx$" stands for something like "is structure-shared in its relevant parts with" (o.c.). As applied to (2), the verb (the selector) shares its morpho-syntactic number with the morpho-syntactic number of the subject (its argument), while the predicative adjective shares its morpho-syntactic number with the number value in the index of the subject. The number agreement in (2b) can be described along the same lines: The AGR|NUMBER value of *on* is unambiguously *singular*, but its INDEX|NUMBER value is underspecified and will be resolved to *plural* if the subject is understood to denote an aggregate.

The purpose of this paper now is to explore how this treatment of the number agreement in copular constructions with a predicative adjective can be extended to the number agreement in copular constructions with a predicate nominal.

[3]Kathol's AGR feature corresponds to Wechsler and Zlatic's CONCORD feature.
[4]Kathol's characterization of (5b) as 'semantic' is misleading, but it is part of the quote.

1.2 Predicate nominals

Predicate nominals canonically show number agreement with the subject, not only in the Romance languages, but also in the Germanic ones, as illustrated for English in (6).[5]

(6) a. His brother is an engineer.
b. His brothers are both engineers.
c. * His brother is engineers.
d. * His brothers are both an engineer.

Mismatches, however, are not excluded. The sentences in (7), for instance, are well-formed.[6]

(7) a. I am best friends with the president of Finland.
b. His brothers are a danger on the road.

The situation is, hence, comparable to that for the predicative adjectives in the Romance languages: In 'normal' circumstances the predicate nominals show morpho-syntactic agreement with the subject, as in (6), but mismatches do not necessarily lead to ill-formedness and suggest that we need a treatment in terms of a more abstract type of agreement. An obvious first choice is Kathol's scheme for 'semantic' agreement in (5b). This, it will be shown, provides an adequate account for some types of mismatch, but not for all.

To substantiate this claim we adopt a usage-based approach. Making use of a treebank which includes a layer of linguistic annotation that identifies subjects, verbs, predicative complements and their respective NUMBER values, we inductively define a typology of mismatches, augmented with frequency data. We then investigate how the different types can be modeled and propose a treatment which subsumes them all.

[5]Predicate nominals do not show gender agreement with the subject.
[6](7a) is quoted from Fillmore et al. (2012, 351).

2 The LASSY treebank

The treebank we will use for the investigation is LASSY. It was constructed in the framework of the STEVIN program (Spyns and Odijk 2013) and is described in Van Noord et al. (2013). This choice admittedly narrows the empirical basis to one language, Dutch, but we believe nonetheless that the analysis will tell us something about number agreement in general, since the Dutch equivalents of (6) and (7) have the same acceptability status as their English counterparts.[7]

(8) a. Zijn broer is een ingenieur.
b. Zijn broers zijn allebei ingenieurs.
c. * Zijn broer is ingenieurs.
d. * Zijn broers zijn allebei een ingenieur.

(9) a. Ik ben beste maatjes met de president van Finland.
b. Zijn broers zijn een gevaar op de weg.

	Contents	# sentence	# word
wr-p-p	Books, brochures, newspapers, reports, periodicals and magazines, proceedings, legal texts, policy documents, surveys, guides and manuals	17,691	281,424
wr-p-e	E-magazines, newsletters, web sites, teletext pages	14,420	232,631
ws-u	Auto cues, news scripts, text for the visually impaired	14,032	184,611
dpc	Dutch Parallel Corpus	11,716	193,029
wikipedia	Dutch Wikipedia pages	7,341	83,360
		65,200	975,055

Table 1: Contents of the LASSY treebank

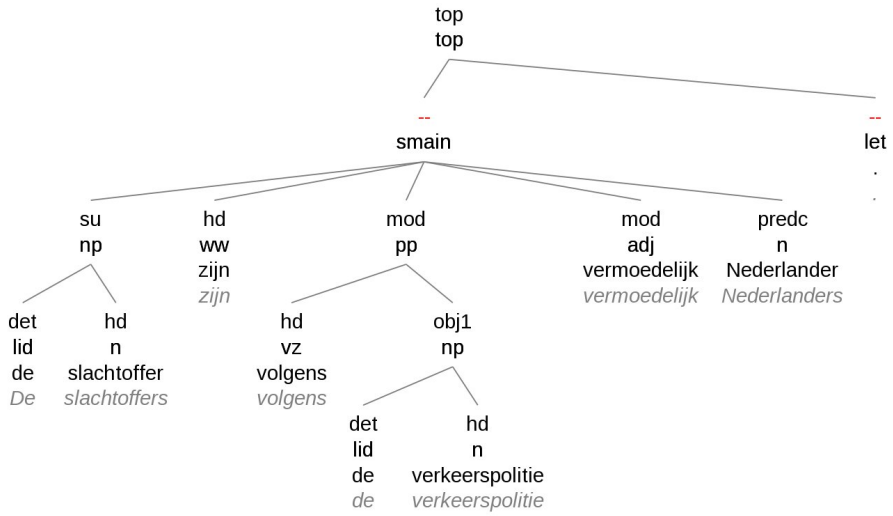
Table 1 provides a survey of the types of texts that the treebank contains and of its size in terms of sentences and words.[8] The texts are divided in sentences and each sentence has a unique identifier, as in (10).

[7] (8d) is well-formed if the indefinite article is omitted. This will be discussed in section 5.2.

[8] These are the numbers for LASSY Small, i.e. the part of the treebank that has been manually checked and, if necessary, corrected. There is also LASSY Large, in which the output of the parser is not manually checked.

(10) De slachtoffers zijn volgens de verkeerspolitie vermoedelijk Nederlanders. [ws-u-e-a-0000000205.p.18.s.2]
'The victims are probably Dutch according to the traffic police.'

Each sentence is assigned a tree that contains information about syntactic categories and dependencies, in accordance with the annotation guidelines in Hoekstra et al. (2003). The tree of (10), for instance, looks as follows.



The word tokens at the bottom of the tree are assigned a lemma, a lexical category, such as adj or n, and a label which defines their role in the phrase, such as h(ea)d or mod(ifier).[9] The phrases are assigned a phrasal category, such as np or pp, and a role. All phrasal nodes have at least two daughters. The daughters of the top node include the sentence-final punctuation and are not assigned a role.

The lexical categories are abbreviations of more detailed part-of-speech tags which contain information about various morpho-syntactic distinctions, in accordance with the annotation guidelines in Van Eynde (2003). These tags are important for our investigation, since they include information about number. More specifically, the nouns and the pronouns have a NUMBER feature, whose value is either singular or plural. For the pronouns, the value may also be underspecified. This is, for instance, the case for the demonstrative *die*, which is singular in (11a) and plural in (11b).[10]

[9]The names of the lexical categories are abbreviations of Dutch terms: "ww" is short for "werkwoord" (verb), "vz" for "voorzetsel" (preposition), and so on.

[10]Notice that we are talking about underspecification of the **morpho-syntactic** number value at this point. Underspecification of the number value in the index will be discussed in section 4.

	Noun	Pronoun	Sum	%
Singular	188,297	25,900	214,197	71.30
Plural	58,458	8,265	66,723	22.21
Underspecified	0	19,486	19,486	6.49
Sum	246,755	53,651	300,406	100.00

Table 2: The (pro)nouns and their number in the LASSY treebank

(11) a. Die komt niet.
'That one does not come.'

b. Die komen niet.
'Those do not come.'

Table 2 provides some quantitative data about the nouns, the pronouns and their morpho-syntactic number in the treebank. It is mainly the figures in the last column that matter in this context, since they provide a base line for measuring the agreement effect. For instance, given that 22.21 % of the (pro)nouns are plural, there is a 22.21 % chance that a random (pro)noun will be plural. If we find a significantly higher percentage in a certain type of context, then this is likely to be a context which —to some quantifiable degree— requires number agreement.

3 Extracting the relevant data

This section describes how we extracted the relevant data from the treebank (3.1) and how we manually checked them for relevance (3.2). The results are summed up in 3.3.

3.1 Querying the treebank

The information which is needed to check the number agreement is extracted by means of queries that are expressed in terms of XPath notation.[11] (12), for in-

[11] See http://www.w3.org/TR/xpath/ for a description of the notation. The queries can be expressed directly in XPath notation or they can be derived automatically from a sample sentence. For this purpose, we use GRETEL, a search tool which automatically parses a given sample sentence, using the ALPINO parser, which allows the user to identify those aspects of the parse that are considered relevant for the search, and which automatically translates the resulting pattern in an XPath query, see http://nederbooms.ccl.kuleuven.be, as well as Augustinus et al. (2012).

stance, retrieves the combinations of a verb, a subject and a predicative complement in which both the subject and the predicative complement have a NUMBER value.[12]

(12) //node[
node[@rel="hd" and @pt="ww"] and
node[@rel="su" and @getal] and
node[@rel="predc" and @getal]] (164 hits)

Since the NUMBER feature is only assigned to lexical categories this query only retrieves combinations in which the subject and the predicative complement are single words. To also retrieve the relevant phrasal categories we need more complex queries in which the subject or the predicative complement contains a head that has the NUMBER feature. (13), for instance, retrieves the combinations in which both are phrasal.

(13) //node[
node[@rel="hd" and @pt="ww"] and
node[@rel="su" and node[@rel="hd" and @getal]] and
node[@rel="predc" and node[@rel="hd" and @getal]]] (1527 hits)

Predictably, there are also combinations in which only the subject is phrasal (129 hits) and in which only the predicative complement is phrasal (1915 hits).[13]

Next, we add specific values for the NUMBER features of the subject and the PREDC. (14), for instance, retrieves the combinations in which both are lexical and singular.

(14) //node[
node[@rel="hd" and @pt="ww"] and
node[@rel="su" and @getal="ev"] and
node[@rel="predc" and @getal="ev"]] (130 hits)

Repeating such queries for the other combinations, we get the data that are displayed in Table 3. The first column specifies whether the subject and the predicative complement are lexical (X) or phrasal (XP); the other columns provide the numbers for combinations with respectively a singular subject, a plural subject

[12]The mutual order does not matter. The names of the morpho-syntactic features and their values are (abbreviations of) Dutch terms: @getal, for instance, stands for number, "ev" for "enkelvoud" (singular) and "mv" for "meervoud" (plural).

[13]The queries do not retrieve coordinate NPs, since they do not have a head daughter.

SU–PREDC	sg-sg	sg-und	sg-pl	pl-sg	pl-und	pl-pl	und-x	Sum
X–X	130	2	8	12	0	7	5	164
XP–X	79	0	11	19	0	18	2	129
X–XP	1640	11	142	53	0	46	23	1915
XP–XP	1272	4	22	137	0	90	2	1527
Sum	3121	17	183	221	0	161	32	3735
%	83.56	0.45	4.90	5.92	0	4.31	0.86	100

Table 3: Result of the queries

and a subject with the underspecified NUMBER value. Since the latter are irrelevant for checking agreement, we do not make finer-grained distinctions in that category: x stands for any of singular, plural or underspecified. Of special interest are the mismatches: There are 183 combinations of a singular subject with a plural predicate nominal, and 221 of a plural subject with a singular predicate nominal. They jointly account for 10.82 % of the combinations.

3.2 Eliminating false and/or irrelevant hits

Manual inspection of the mismatches reveals that not all of them count as genuine mismatches. This is due to a number of reasons which we will partition in four groups.

The first reason concerns the fact that the queries do not differentiate between subject-oriented and object-oriented predicative complements. As a consequence, the sentence in (15) ends up among the mismatches, since it combines a plural subject with a singular predicate nominal, but it does not count as a genuine mismatch, since object-oriented predicative complements are expected to show number agreement with the direct object, rather than with the subject.

(15) Alle personages noemen hem “Ron” ... [wr-p-e-i-0000004258.p.3.s.1]
‘All characters call him “Ron”...’

The second reason concerns the existence of disfluencies, as in (16).

(16) Ook voor ... de politieke discussies in de samenleving zijn ze *groot belang*. [wr-p-p-j-0000000013.p.39.s.2]
‘They are also (of) great importance for ... public political discussions.’

The italicized predicate nominal does not show number agreement with the plural subject, but this is no surprise, since it should be introduced by the preposition *van* 'of'. If the preposition is added, the clause no longer counts as a mismatch, since prepositional PREDCs do not show number agreement.

The third reason concerns annotation errors. The predicative complement in (17), for instance, is analyzed as a plural noun, but in the given clause it is an adjective, and, hence, exempt from number agreement. Similarly, the predicate nominal in (18) is labeled as plural, but is in fact singular, just like the subject.

(17) De ambtenaar is lui. [dpc-vla-001161-nl-sen.p.117.s.2]
'The civil servant is lazy.'

(18) De hoofdstad van Wallonië is Namen. [wiki-135.p.36.s.1]
'The capital of Wallonia is Namur.'

The fourth reason concerns the existence of a certain amount of friction between the purpose of this investigation and the choices that have been made in the annotation guidelines. The choice, for instance, to restrict the use of the underspecified NUMBER value to the pronouns forces the parser to assign a specific value to all non-pronominal NPs, also when the underspecified value would have been more appropriate. Some relevant instances are the italicized subjects in (19) and (20).

(19) Vooral in de katoensector is *de VS* de grootste en meest schadelijke subsidieverstrekker. [wr-p-p-j-0000000001.p.108.s.1.2.]
'Especially in the cotton sector the US is the largest and most noxious provider of subsidies.'

(20) Zijn overzicht van de wetten beslaat 14 koppen, maar *een aantal daar van* zijn alleen maar definities. [wr-p-e-i-0000041235.p.1.s.124]
'His survey of the laws comprises 14 headings, but a number of those are just definitions.'

De VS, which is short for *de Verenigde Staten* 'the United States', is treated as plural in the treebank, and the partitive subject in (20) as singular, since it is headed by the singular noun *aantal* 'number', but given that they equally combine with singular and plural verbs, as shown in (21), it would make more sense to treat their morpho-syntactic NUMBER as underspecified.[14]

[14] In this respect, they resemble the demonstrative *die*, see (11).

(21) a. Vooral in de katoensector is/zijn de VS niet meer competitief.
'Especially in the cotton sector the US is/are no longer competitive.'

b. Zijn overzicht van de wetten beslaat 14 koppen, maar een aantal daar van is/zijn nu al verouderd.
'His survey of the laws comprises 14 headings, but a number of those is/are already obsolete.'

In that case the combinations in (19) do not qualify as mismatches, but rather as combinations of a predicate nominal with a specific NUMBER value and a subject with an underspecified NUMBER value.

Another point of friction concerns the identification of the head in nominal phrases. The head of the italicized predicate nominal in (22), for instance, is identified with the plural common noun, yielding a mismatch with the singular subject.

(22) Bij een vrouw is de grens *veertien glazen*. [wr-p-p-c-0000000048.txt-341]
'For a woman the limit is fourteen glasses.'

Notice, though, that the same nominal, when used in subject position, is compatible with both plural and singular verbs.

(23) a. Veertien glazen zijn tijdens de verhuis gebroken.
'Fourteen glasses were broken during the move.'

b. Veertien glazen is ruim voldoende.
'Fourteen glasses is amply sufficient.'

The difference corresponds with a semantic distinction: In (23a) the VP says something about the glasses themselves, but in (23b) it says something about their number. This provides good evidence for treating the numeral as the head of the subject in (23b). Evidence for treating it as singular is provided by (24).

(24) Veertien is/*zijn mijn geluksgetal.
'Fourteen is/*are my lucky number.'

Notice also that Dutch numerals show the same inflectional variation as common nouns: They can take a diminutive affix, as in *een tientje* 'a tenner', a plural affix, as in *honderden* 'hundreds' or a dative affix, as in *met z'n vieren* 'with four people', see Van Eynde (2006). Returning now to the combination in (22), since the italicized predicate nominal mainly provides information about the number of glasses, it makes good sense to identify its NUMBER value with that of the

numeral, and in that case it no longer qualifies as a mismatch, but rather as the combination of a singular predicate nominal with a singular subject.

A third and final source of friction concerns the treatment of autoreferential nominals, as in (25).

(25) Het thema dit jaar is "*Steden*". [wr-p-e-c-0000000004.p.15.s.6]
'The theme this year is "Cities".'

While the italicized predicative complement undeniably contains a plural nominal and even nothing but a plural nominal, it does not qualify as plural for the purpose of agreement. Notice, for instance, that it requires a singular verb when it occurs in subject position, as in (26).

(26) "Steden" lijkt/*lijken me wel een geschikte titel voor dit boek.
"'Cities" seems/*seem an appropriate title for this book.'

Apparently, the autoreferential use involves a neutralization of the number distinction to singular. In fact, it also involves the neutralization of the categorial distinction to nominal, as shown in (27), where the autoreferential PP occurs in a position which is normally reserved for NPs.

(27) Ze heeft "Uit Afrika" op twee maand tijd geschreven.
'She wrote "Out of Africa" in two months' time.'

It is, hence, natural to treat the autoreferential predicative complement in (25) as a singular nominal, and in that case it no longer qualifies as a mismatch.

3.3 Result

The result of the manual inspection of the automatically retrieved data is summed up in Table 4. The numbers in the first row are identical to those in the bottom line of Table 3. The hits for object-oriented predicative complements are subtracted for all types of combinations. Disfluencies and annotation errors have only been identified for the combinations with a putative mismatch. The annotation errors are partitioned in two types. The first concerns those which, when corrected, no longer belong to the relevant sample space, for instance, because the predicative complement is not nominal, as in (17). The second type concerns those which, when corrected, belong to another subclass of the sample space, for instance, because the predicate nominal is not plural but singular, as in (18). Mismatches which are due to friction between the annotation guidelines and the objectives of

SU−PREDC	sg-sg	sg-und	sg-pl	pl-sg	pl-pl	und-x	Sum
	3121	17	183	221	161	32	3735
Obj-Or PredC	-26	0	-3	-21	-2	-2	-54
Disfluencies			0	-1			-1
Annot error T1			-7	-12			-19
Annot error T2	3		3	-6			0
Friction	15		-19	-1		5	0
Result	3113	17	157	180	159	35	3661
%	85.03	0.46	4.29	4.92	4.73	0.96	100

Table 4: Matches and mismatches in the LASSY treebank

the present investigation are also subtracted and added to the columns where they belong. As a result of these adjustments, the amount of mismatches has been reduced from 10.82 % to 9.21 %.

Zooming in on combinations with a specific number value for the subject, we observe a striking asymmetry: While the proportion of mismatches in clauses with a singular subject (3287) is no more than 4.78 %, the proportion in clauses with a plural subject (339) is a whopping 53.10 % ! To measure the agreement effect we should compare these to the average frequency of plurals and singulars, respectively, as given in Table 2. The former is 22.21 %, which is well above 4.78 %. The latter is 71.30 %, and while this is obviously more than 53.10 %, it is clear that the agreement effect is less strong.

4 A typology of mismatches

So far, we have made a distinction between two types of mismatch: singular subjects vs. plural predicate nominals and plural subjects vs. singular predicate nominals. Orthogonal to this distinction we add another one based on the morpho-syntactic number of the verb. As shown in Table 5, nearly all of the mismatches in the LASSY treebank occur in sentences with a plural verb.

We now take a close look at the four types of mismatch. In a first go, we discuss the types one by one, investigating whether they can be modeled in terms of the agreement schemata that are proposed in Kathol (1999). If not, we present an alternative. In the fifth paragraph, we pull the various strands together and propose a unified account that also includes the number agreement with the verb.

SU–PREDC	sg-pl	pl-sg	Sum	%
Plural verb	155	174	329	97.63
Singular verb	2	6	8	2.37
Sum	157	180	337	100

Table 5: Four types of mismatches

4.1 Singular subject vs. plural verb and plural PREDC

Of the 155 instances of this type, no less than 151 concern combinations in which the subject is the impersonal neuter pronoun *het* 'it' or one of the demonstrative neuter pronouns *dat* 'that' and *dit* 'this'. Some examples are given in (28–29).[15]

(28) Het worden spannende maanden. [dpc-vhs-000759-nl-sen-p.28.s.1]
'It'll be exciting months.'

(29) Dit zijn uiterst verontrustende berichten. [dpc-bal-001239-nl-sen-p.60.s.1]
'These are very worrying messages.'

These pronouns are also compatible with a singular predicate nominal, as shown in (30). In that case the verb is singular too.

(30) a. Het wordt/*worden een spannende maand.
'It'll be an exciting month.'
b. Dit is/*zijn een uiterst verontrustend bericht.
'This is a very worrying message.'

If the predicative complement is not nominal, the pronouns are only compatible with a singular verb.

(31) a. Het wordt/*worden ongemeen spannend.
'It'll be very exciting.'
b. Dit is/*zijn uiterst verontrustend.
'This is very worrying.'

In this respect, they differ from the nominals with an underspecified morpho-syntactic NUMBER value, such as the demonstrative *die*, the proper noun *VS* and

[15]The German equivalents *es, das* and *dies* are also compatible with plural verbs and predicate nominals. The English equivalents *it, that* and *this*, by contrast, are not.

the partitive subjects, see (11) and (21). This suggests that *het, dat* and *dit* are morpho-syntactically singular, but that their INDEX|NUMBER value is underspecified. In that respect, they resemble the French impersonal pronoun *on* 'one' in (2b).

The four remaining instances of this type have a non-pronominal subject. Three of them are listed in (32–34).[16]

(32) De kleding die ze droegen waren vermoedelijk dierenvellen. [wr-p-e-i-0000050381.p.1.s.16]
'The clothing which they wore were probably animal hides.'

(33) EVISTA zijn gele, ovaalvormige tabletten met de werkzame stof raloxifenehydrochloride (60 mg). [dpc-eli-000941-nl-sen.p.8.s.1]
'EVISTA are yellow, oval shaped tablets with ...'

(34) Een kind kan zien dat het trio van de 'As van het kwaad' toevallig ook de vijanden van Israël zijn. [wr-p-p-i-0000000172.p.3.s.10]
'Even a child can see that the trio of the 'Axis of evil' are incidentally also the enemies of Israel.'

The subjects in (32) and (33) are headed by a noun that lacks a plural counterpart: *kleding* is a mass noun, and *EVISTA* is the name of a product. The one in (34) does not lack a plural counterpart, but it denotes a collective, which implies that it can be interpreted in two ways: as semantically singular if it stands for the collective, and as semantically plural if it stands for the members of the collective.[17] The subjects in (32–34) are also compatible with a singular predicate nominal. In that case the verb is singular too.

(35) a. De kleding die ze droegen was/*waren een simpel dierenvel.
'The clothing which they wore was/*were a simple animal hide.'

b. EVISTA is/*zijn een geel, ovaalvormig tablet.
'EVISTA is/*are a yellow, oval shaped tablet.'

c. Het trio van de 'As van het kwaad' is/*zijn een gevaarlijke alliantie.
'The trio of the 'Axis of evil' is/*are a dangerous alliance.'

Moreover, if the predicative complement is not nominal, the verb must be singular.

[16]The fourth one is nearly identical to (33); the only difference concerns the subject: *OPTRUMA* instead of *EVISTA*.

[17]The verb *zijn* 'are' in (34) is equivalent to the singular *bestaat uit* 'consists of'.

(36) a. De kleding die ze droegen was/*waren versleten.
'The clothing they wore was/*were worn out.'
b. EVISTA is/*zijn geel en ovaalvormig.
'EVISTA is/*are yellow and oval shaped.'
c. Het trio van de 'As van het kwaad' is/*zijn niet meer gevaarlijk.
'The trio of the 'Axis of evil' is/*are no longer dangerous.'

This shows that these subjects are similar to the pronouns in (28–29), in the sense that they are morpho-syntactically singular, while their INDEX|NUMBER value is underspecified.[18]

To model the agreement between the subject and the predicate nominal, we can use Kathol's scheme for semantic agreement in (5b), repeated in (37).

(37) AGR(selector) $\approx$ INDEX(argument)

Assuming that the predicate nominal is the selector and that number is the property to be shared, this scheme identifies the morpho-syntactic NUMBER value of the predicate nominal with the INDEX|NUMBER value of the subject. If the latter is underspecified, as in the examples above, the application of the scheme has the effect of resolving the underspecification, setting it to plural in combinations with a plural subject, as in (28–29) and (32–34), and to singular in combinations with a singular subject, as in (30) and (35).

Notice that (37) can also be used to model the agreement between subject and verb. Assuming that the verb is the selector, the INDEX|NUMBER value of the subject is set to plural in (28–29) and (32–34), and to singular in (30) and (35). The interaction of this constraint on agreement with the one on the agreement between subject and predicate nominal accounts for the ill-formedness of the starred variants in (30) and (35).

4.2 Plural subject and plural verb vs. singular PREDC

With 174 instances, this is the most common type of mismatch. It comes in two subtypes. The first includes combinations in sentences with a **collective** interpretation. An example from English was given in (7b) and is repeated in (38).

[18] Further evidence for treating *trio* as morpho-syntactically singular is provided by the presence of the definite article *het*. This article only combines with singular neuter nouns; plural nouns and non-neuter singular nouns require the article *de*.

(38) His brothers are a danger on the road.

Corresponding examples from the LASSY treebank are given in (39–40).

(39) De Tsjetsjeense strijders zijn een relatief kleine groep. [ws-u-e-a-0000000244.p.3.s.4]
'The Chechen warriors are a relatively small group.'

(40) Politieke tegenstellingen zijn een wezenskenmerk van elke democratie. [dpc-kok-001320-nl-sen.p.6.s.2]
'Political contrasts are a defining characteristic of every democracy.'

(39) does not mean that every single Chechen warrior forms a small group, but rather that the collective of them does. Similarly, (40) does not mean that every single political contrast is a defining characteristic of democracy, but rather that the existence of political contrasts in general is a characteristic of democracy. The same subjects can be used in combination with a plural predicate nominal, as in (41).

(41) a. De Tsjetsjeense strijders zijn gevaarlijke terroristen.
'The Chechen warriors are dangerous terrorists.'

b. Politieke tegenstellingen zijn vaak obstakels in het zoeken naar een zinnige oplossing.
'Political contrasts are often obstacles in the search for a sensible solution.'

This suggests that they have an underspecified INDEX|NUMBER value which is resolved to singular in (39–40) and to plural in (41). As such, this subtype fits the mould of Kathol's scheme for semantic agreement: (37) identifies the INDEX|NUMBER value of the subject with the morpho-syntactic number of the predicate nominal.

What does not work, though, is the interaction with the constraint on subject-verb agreement: If the INDEX|NUMBER value of the subject is identified with the morpho-syntactic number of the verb, then it is set to plural in (39–40), contradicting the result of the agreement with the predicate nominal which sets it to singular.

The second subtype includes combinations in sentences with a **distributive** interpretation. Some examples from the LASSY treebank are given in (42–43).

(42) Beide aftredende bestuurders blijven wel aandeelhouder. [wr-p-e-i-0000049645.p.1.s.68.2]
'Both resigning directors do remain shareholder.'

(43) Veel Vandersteen-fans zijn daarom meer liefhebber van zijn strips uit de jaren ’40 en ’50. [wiki-9843.p.18.s.4]
‘For that reason many fans of Vandersteen prefer his comics from the forties and the fifties.’

(42) is about two resigning directors who both remain shareholders, and (43) is about a multitude of fans who all prefer Vandersteen’s comics from the fourties and the fifties. The predicate nominals in these combinations are also compatible with a singular subject, as in (44).

(44) a. De aftredende bestuurder blijft wel aandeelhouder.
‘The resigning director does remain a shareholder.’
b. Deze Vandersteen-fan is liefhebber van zijn strips uit de jaren ’50.
‘This fan of Vandersteen is in favor of his comics from the fifties.’

This suggests that they have an underspecified INDEX|NUMBER value, which is resolved to plural in (42–43) and to singular in (44).

To model the agreement the scheme in (37) is inappropriate: If the subject’s INDEX|NUMBER value is required to be identical with the morpho-syntactic NUMBER value of the predicate nominal, then the quantified subjects in (42) and (43) are assigned a singular index, which is clearly at odds with their meaning. Moreover, it is also at odds with the application of (37) to the agreement between subject and verb.

A scheme that better fits the data is one which requires sharing between the INDEX|NUMBER value of the predicate nominal and the morpho-syntactic number of the subject, as in (45).

(45) INDEX(selector) $\approx$ AGR(argument)

This allows the singular predicate nominals with their underspecified index to combine both with singular and plural subjects, resolving the underspecification in the process.

Notice that the underspecification only holds for the singular forms of the predicate nominals. Their plural counterparts are only compatible with a plural subject.

(46) a. De aftredende bestuurders blijven wel aandeelhouders.
‘The resigning directors do remain shareholders.’
b. Deze Vandersteen-fans zijn liefhebbers van zijn strips uit de jaren ’50.
‘These fans of Vandersteen prefer his comics from the fifties.’

(47) a. * De aftredende bestuurder blijft wel aandeelhouders.
b. * Deze Vandersteen-fan is liefhebbers van zijn strips uit de jaren '50.

This shows that the INDEX|NUMBER value of the plural forms is *plural* rather than underspecified. In combination with (45) this accounts for the ungrammaticality of (47).

Taking stock, Kathol's scheme for semantic agreement can cope with the mismatches in combinations with a collective interpretation, but not with those in combinations with a distributive interpretation. Besides, it has problems with the interaction of the constraints on subject-verb agreement and subject-predicate nominal agreement.

4.3 Singular subject and singular verb vs. plural PREDC

This is the type of mismatch that is exemplified by the English sentence in (7a), repeated in (48).

(48) I am best friends with the president of Finland.

It is a rare phenomenon in Dutch. The LASSY treebank contains only two instances. They are listed in (49–50).

(49) Goud blijft de belangrijkste financiële activa van bijna alle centrale banken. [wr-p-e-i-0000032165.p.5.s.255]
'Gold remains the most important financial activa of nearly all central banks.'

(50) Anders is het geen domotica. [dpc-rou-000479-nl-sen.p.10.s.8]
'Otherwise it is not domotics.'

Activa and *domotica* are nouns that lack a singular counterpart. Unsurprisingly, they are also compatible with a plural subject, as in (51).

(51) a. Goudreserves blijven de belangrijkste financiële activa van bijna alle centrale banken.
'Gold reserves remain the most important financial activa of nearly all central banks.'
b. Stofzuigers waren de domotica van die tijd.
'Vacuum cleaners were the domotics of that time.'

The data in (49–50) provide further evidence against (37), since the latter erroneously requires the subjects to have a plural index. The scheme in (45) is more appropriate: Assuming that the pluralia tantum have an underspecified INDEX|NUMBER value which is identified with the morpho-syntactic number of the subject, it is set to singular in (49–50) and to plural in (51).

4.4 Plural subject vs. singular verb and singular PREDC

This type of mismatch is uncommon too. LASSY contains only six instances. Two are listed in (52–53).

(52) De Vulcans is een ras van zeer intelligente mensachtigen, die logica als de basis voor iedere beslissing zien. [wr-p-e-i-0000027197.p.3.s.155.2]
'The Vulcans is a race of very intelligent humanoids, who see logic as the basis of every decision.'

(53) De minderheden was echter het zwakke punt van de jonge staat. [wr-p-e-i-0000051928.p.1.s.41]
'The minorities was, however, the weak point of the young nation.'

These sentences have a collective interpretation and can be analyzed along the same lines as those in (39). The only difference concerns the form of the verb, but that difference is not crucial, as demonstrated by the fact that the sentences in (54) are no less well-formed than those in (52).

(54) a. De Vulcans zijn een ras van zeer intelligente mensachtigen.
'The Vulcans are a race of very intelligent humanoids.'

b. De minderheden waren echter het zwakke punt van de jonge staat.
'The minorities were, however, the weak point of the young nation.'

Moreover, if the predicative complement is not nominal, the verb must be plural, as shown in (55).

(55) a. De Vulcans zijn/*is zeer intelligent.
'The Vulcans are/*is very intelligent.'

b. De minderheden waren/*was echter zwak en verdeeld.
'The minorities were/*was, however, weak and divided.'

In that respect, these subjects contrast with *de VS*, which is also compatible with a singular verb in this context. This shows that the subjects in (52–55) are morpho-syntactically plural, rather than underspecified. Their index, however, is underspecified: Its NUMBER value is resolved to singular if the predicate nominal is

singular, as in the examples above, and to plural if the predicate nominal is plural, as in (56).

(56) a. De Vulcans zijn zeer intelligente mensachtigen.
'The Vulcans are very intelligent humanoids.'

b. De minderheden waren echter gezworen vijanden van de jonge staat.
'The minorities were, however, sworn enemies of the young nation.'

This fits the mould of Kathol's scheme for semantic agreement in (37).

4.5 A unified account

At this point, we have two schemata for modeling the agreement between predicate nominals and subjects. There is Kathol's scheme for semantic agreement, repeated in (57), and the alternative scheme, repeated in (58).

(57) AGR(selector) $\approx$ INDEX(argument)

(58) INDEX(selector) $\approx$ AGR(argument)

The former deals with the first and the fourth type of mismatch, as well as with the collective subtype of the second; the latter deals with the third type of mismatch and with the distributive subtype of the second. This obviously raises the suspicion that we are missing a generalization.

To repair this we develop a treatment which is based on the scheme in (59).

(59) index agreement: INDEX(argument) $\approx$ INDEX(argument)

In contrast to the schemata in (57) and (58), this scheme requires index agreement pure and simple, rather than an asymmetric kind of agreement between the AGR value of one term and the INDEX value of the other term. Besides, it does away with the asymmetry in the relation between predicative complement and subject: instead of treating one as the selector and the other as its argument, they are treated as co-arguments. Empirical evidence for the assumption that subject and predicative complement are co-arguments in copular constructions is provided in Van Eynde (2009) and Van Eynde (2012).

Adopting this style of analysis the number agreement can be modelled in terms of a constraint on the predicate selecting lexemes, as in (60).[19]

[19]Nearly all of the predicate selecting lexemes are verbs. The only exceptions (in Dutch) are the prepositions *met* 'with' and *zonder* 'without' in absolutive constructions. Since they are only

(60)
$$\left[\begin{array}{l} \textit{predicate-selecting-lexeme} \\ \text{ARG-ST} \quad \langle \text{NP}_{\boxed{1}}, \text{NP}_{\boxed{2}} \rangle \\ \\ \ldots \mid \text{CONTENT} \mid \text{NUCLEUS} \left[\begin{array}{l} \text{THEME} \quad \boxed{1}\left[\begin{array}{l} \textit{index} \\ \text{NUMBER} \quad \boxed{3}\,\textit{number} \end{array}\right] \\ \\ \text{ATTRIBUTE} \quad \boxed{2}\left[\begin{array}{l} \textit{index} \\ \text{NUMBER} \quad \boxed{3} \end{array}\right] \end{array}\right] \end{array}\right]$$

The ARG-ST value lists the constituents which are selected by the lexeme. In this case, it includes two NPs. Their indices are identified with the values of the THEME feature and the ATTRIBUTE feature respectively. This sharing models the assignment of semantic roles. The crucial part concerns the NUMBER values in the indices: They are required to be identical.

Notice that the constraint only concerns the NUMBER values. It does not mention the other features which figure in the indices, such as PERSON and GENDER, since these do not need to be identical in copular constructions. Likewise, it does not concern the morpho-syntactic NUMBER values. As a consequence, the constraint in (60) allows morpho-syntactic number mismatches. More specifically, it allows them if the INDEX|NUMBER value of a nominal is different from its AGR|NUMBER value. In other words, mismatches are attributed to discrepancies between the INDEX|NUMBER value and the AGR|NUMBER value of either the subject or the predicate nominal.

Another salient property of the agreement constraint in (60) is that it concerns a property of lexemes, not of words. It, hence, generalizes over the various forms that verbs can take, both the finite and nonfinite ones. Assuming that lexemes are related to words by means of inflectional lexical rules, as in Sag et al. (2003), we need at least as many inflectional rules as there are distinct verb forms. The number differs from language to language. For Dutch, we need at least one rule for the plural forms and two for the singular forms. The former is spelled out in (61).

compatible with adjectival and prepositional predicative complements, they are irrelevant for the treatment of number agreement.

(61)
$$\left[\begin{array}{l}\textsc{input}\ \left\langle \boxed{1}, \textit{verb-lxm} \right\rangle \\ \textsc{output}\ \left\langle \mathrm{F}_{pl}\left(\boxed{1}\right), \left[\begin{array}{l}\textit{word} \\ \textsc{vform}\ \textit{finite} \\ \textsc{subj}\ \left\langle \textsc{np}\left[\begin{array}{l}\textsc{case}\ \textit{nominative} \\ \textsc{agr}\,|\,\textsc{number}\ \textit{plural}\end{array}\right]\right\rangle\end{array}\right]\right\rangle\end{array}\right]$$

The input and the output are both pairs that consist of a morpho-phonological form and a structured set of syntactic and semantic properties. F_{pl} is a function which maps a verbal stem onto its plural form. For most verbs, this involves the addition of the suffix *-en*, as in *worden* 'become'. The resulting form is finite and requires its subject to be nominative and morpho-syntactically plural.

For the singular forms we need separate rules for the forms without suffix and the forms with the suffix *-t*, as in *word* and *wordt*. Both yield a finite form that requires its subject to be nominative and morpho-syntactically singular.[20]

Together, these lexical rules model the agreement between subject and verb in the vast majority of cases. An exception must be made, though, for combinations which display the kind of mismatch which is exemplified in (62).

(62) a. Het worden spannende maanden.
'It'll be exciting months.'

b. De kleding die ze droegen waren vermoedelijk dierenvellen.
'The clothing they wore were probably animal hides.'

These are not allowed by a grammar which only disposes of (61) to model the plural forms. To accommodate them we add the lexical rule in (63).

[20]What differentiates the singular forms is a rather intricate mixture of person, mood, tense and position (V1, V2 or V-final) distinctions.

(63)
$$\left[\begin{array}{l}\textsc{input}\ \left\langle \boxed{1}, \left[\begin{array}{l}\textit{verb-lxm}\\ \textsc{arg-st}\ \left\langle \textsc{np}_{\boxed{2}}, \textsc{np}_{\boxed{3}}\right\rangle\\ \textsc{content}\,|\,\textsc{nucleus}\ \left[\begin{array}{ll}\textsc{theme} & \boxed{2}\ \textit{index}\\ \textsc{attribute} & \boxed{3}\ \textit{index}\end{array}\right]\end{array}\right]\right\rangle\\ \textsc{output}\ \left\langle \mathrm{F}_{pl}\left(\boxed{1}\right), \left[\begin{array}{l}\textit{word}\\ \textsc{vform}\ \ \textit{finite}\\ \textsc{subj}\ \left\langle \textsc{np}\left[\begin{array}{ll}\textsc{case} & \textit{nominative}\\ \textsc{agr}\,|\,\textsc{number} & \textit{singular}\\ \textsc{index} & \boxed{2}\left[\textsc{number}\ \ \textit{plural}\right]\end{array}\right]\right\rangle\end{array}\right]\right\rangle\end{array}\right]$$

The morpho-phonological part of the rule is the same as in (61), but its range of application is much smaller: It is restricted to the lexemes which are subsumed by the constraint in (60). The plural forms of such lexemes combine with subjects that are morpho-syntactically singular, but whose INDEX|NUMBER is plural. Given the general constraint on agreement between subject and predicate nominal in (60), this implies that the INDEX|NUMBER value of the predicate nominal is plural as well.

Since the output of rule (63) explicitly requires the subject to have conflicting values for morpho-syntactic number and INDEX|NUMBER, the resulting verbs are only compatible with singular subjects that allow their INDEX|NUMBER to be resolved to plural. This is a property which only some of the nominals have. They include the pronouns *het, dat* and *dit*, singularia tantum, such as *kleding*, and collective nouns, such as *trio*.

The same remarks apply m.m. to the instances of the fourth type of mismatch, as exemplified in (64).

(64) a. De Vulcans is een ras van zeer intelligente mensachtigen.
'The Vulcans is a race of very intelligent humanoids.'

b. De minderheden was echter het zwakke punt van de jonge staat.
'The minorities was, however, the weak point of the young nation.'

These are not allowed by the canonical lexical rule(s) for the singular verb forms. To accommodate them we need rules which allow the singular verb forms to combine with a subject that is morpho-syntactically plural but that has a singular index. These rules look similar to the one in (63).

Summing up, we now have a single constraint for modelling the agreement between subjects and predicate nominals. It is defined in terms of a constraint on the predicate selecting lexemes, see (60). Besides, we have modeled the agreement between subject and verb in terms of inflectional lexical rules which allow the individual verb forms to put constraints on the morpho-syntactic number of their subject, as in (61). To allow for mismatches we have added extra lexical rules with a more restricted range of application, as in (63).

5 The relation between AGR|NUMBER and INDEX|NUMBER

A potential problem for the constraint in (60) is that it might be too permissive. If there are no constraints on the relation between the AGR|NUMBER and the INDEX|NUMBER of the nominals, then the grammar allows any of the combinations in (65) and (66), including those which are ill-formed.

(65) a. Hij is beste maatjes met de president van Finland.
b. Zijn broer is een ingenieur.
c. * Zijn broer is ingenieurs.

(66) a. Zijn broers zijn een gevaar op de weg.
b. Zijn broers zijn allebei ingenieurs
c. * Zijn broers zijn allebei een ingenieur.

It is clear then that we need some constraints on that relation. For the plural predicate nominals the relevant constraint is fairly simple: Their INDEX|NUMBER is identical to their morpho-syntactic NUMBER value, except in the case of pluralia tantum, such as *activa*, and nouns with a reciprocal meaning, such as *vrienden* 'friends'.[21] They have the underspecified value for INDEX|NUMBER and are, hence, compatible with singular and plural subjects alike.

For the singular predicate nominals the issue is more complex. In combinations with a morpho-syntactically plural subject, the constraint on INDEX|NUMBER agreement can be satisfied in one of two ways. One possibility is that the predicate nominal's INDEX|NUMBER value is identical to its morpho-syntactic NUMBER value. Given the constraint on agreement this implies that the subject is assigned a singular index too, so that the combination has a collective interpretation. The other possibility is that the predicate nominal has an underspecified

[21] The treebank contains some examples of the former (see 4.3), but none of the latter. Further investigation is needed to identify the nouns which belong to this class.

INDEX|NUMBER value, which depending on the INDEX|NUMBER value of the subject is resolved to either singular or plural. In the latter case, the resulting interpretation is of the distributive type.

Both possibilities will be exemplified and discussed, first the combinations with collective interpretation (section 5.1) and then those with a distributive interpretation (section 5.2). Since not all sentences neatly fall in one of these two classes we also discuss some sentences for which both interpretations are equally plausible (section 5.3).

5.1 Triggers of a collective interpretation

In combinations with a collective interpretation a singular predicate nominal also has a singular index. Given the constraint on agreement in (60) this implies that the subject has a singular index too, also if it is morpho-syntactically plural. Some examples were already discussed in section 4.2, see (39–40). The purpose of this section is to identify a number of factors which invariably trigger a collective interpretation.

5.1.1 Inherently collective PREDCs

Predicate nominals which are headed by an inherently collective noun, such as *groep* 'group' in (39), have a singular index and, hence, trigger a collective interpretation. Other examples of this kind concern combinations with *volk* 'people', *groepering* 'faction', *verzameling* 'set', *meerderheid* 'majority' and *brassband*. Some are listed in (67–68).

(67) Is het omdat wij een volk van bierdrinkers zijn dat Belgische vorsers zich zo frequent - en met succes - over leverziekten buigen? [wr-p-p-i-0000000100.p.1.s.1]
'Is it because we are a people of beer drinkers that the Belgian scientists so often - and successfully - focus on liver diseases?

(68) Tijdens de Derde Republiek waren de monarchisten de reactionaire groepering van die tijd. [wr-p-e-i-0000041531.p.7.s.8]
'In the Third Republic the monarchists were the reactionary faction of that period.'

(67) does not mean that each of us is a beer drinking people, but rather that we collectively are such a people. In all of the relevant instances the predicate nom-

inals are introduced by a determiner. It may be definite or indefinite, but it may not be omitted.

5.1.2 Predicate nominals with a unique referent

Another group of singular predicate nominals which trigger a collective interpretation are those with a unique referent. They are headed by a singular count noun and must be introduced by a definite determiner. Besides, they often contain an overt uniqueness marker, such as the adjective *enige* 'only' in (69) and the superlative *belangrijkste* 'most important' in (70).

(69) ... omdat in de plannen de roltrappen de enige vluchtweg uit de ondergrondse zijn. [ws-u-e-a-0000000200.p.15.s.2]
'... because the mobile stairs are the only way out from the underground.'

(70) De Leien (Frankrijklei, Italiëlei, Amerikalei, Britselei) zijn de belangrijkste verkeersader binnen Antwerpen. [wiki-11.p.54.s.1]
'The Leien (...) are the most important traffic artery in Antwerp.'

In (70) there is one trajectory that is claimed to be the most important one in Antwerp and that trajectory is identified with the 'Leien' as a whole. If one wants to claim that each one of the four 'Leien' belongs to the most important trajectories of the city, one must use the plural counterpart of the predicate nominal, i.e. *de belangrijkste verkeersaders*.

Other markers of uniqueness are the appositive proper noun in (71) and the emphatic stress on the determiner in (72).

(71) De kernen Heukelom en Montenaken werden de gemeente Vroenhoven en ... [wr-p-e-i-0000050381.p.1.s.148]
'The nuclei Heukelom and Montenaken became the town Vroenhoven and ...

(72) ze zijn sinds jaar en dag dé twistappel tussen Pakistan en India. [wr-p-p-i-0000000259.p.2.s.3]
'they have been THE topic of disagreement between Pakistan and India.'

The examples in (69–72) neatly show the correlation between the assignment of a singular index to the predicate nominal and the asssignment of a collective interpretation to the clause as a whole: It is because the predicate nominals have a unique referent that their index is singular, and it is because of the constraint on the sharing of the INDEX|NUMBER values that the subject has a singular index

as well, in spite of the fact that its morpho-syntactic NUMBER is unambiguously plural.

5.1.3 Topicalized PREDCs

A third group of predicate nominals which trigger a collective interpretation are the topicalized ones, as in (73–74).

(73) Het hoogtepunt in haar sportcarrière waren de vier gouden medailles die ze won bij de Olympische Spelen van 1948. [ws-u-e-a-0000000028.p.22.s.5]
'The climax in her sports carreer were the four gold medals which she won at the 1948 Olympics.'

(74) Een heel specifiek Brussels fenomeen zijn de 22 gemeenschapscentra, die de lokale draaischijf vormen van het Vlaamse sociale en culturele leven. [dpc-vla-001171-nl-sen.p.38.s.1]
'A peculiar Brussels phenomenon are the 22 community centers which form the local pivot of the Flemish social and cultural life.'

These sentences clearly have a collective interpretation: (73) is about one apical moment in her career, not about four such moments, and (74) is about one phenomenon which is claimed to be typical of Brussels, and not about as many phenomena as there are community centers, i.e. 22.

Topicalized predicate nominals may also be bare, as in (75–76).[22]

(75) Grote winnaar bij de verkiezingen van 18 mei 2003 waren de socialisten. [wr-p-e-h-0000000051.p.249.s.1]
'The socialists were the great winners of the elections of May 18, 2003.'

(76) Onzin zijn ook de verhalen dat hij twee buitenechtelijke Londense zoons zou hebben. [ws-u-e-a-0000000037.p.4.s.6]
'Equally nonsensical are the rumors that he supposedly has two extramarital sons in London.'

Also these sentences have a collective interpretation. It is, for instance, not every individual socialist that was the great winner of the 2003 elections, but rather the socialists as a collective.

The reason why sentences with a topicalized singular PREDC have a collective interpretation is probably due to a correlation between linear order and scope:

[22] If they occur in the canonical position in the Mittelfeld, these predicate nominals require a definite article, unless they are headed by a mass noun.

Since the singular PREDC precedes the plural subject, it also outscopes (any quantifiers in) that subject.

5.1.4 Summing up

Singular predicate nominals have a singular index, if they are headed by an inherently collective noun, if they have a unique referent or if they are topicalized. It is sufficient that one of these conditions is fulfilled in order to trigger a collective interpretation, but in individual sentences it may happen that more than one is fulfilled at the same time, as in (77) and (78).

(77) Onder deze laatsten waren de Grieken de grootste groep. [wr-p-e-i-0000000001.p.7.s.4]
'Among these the Greeks were the largest group.'

(78) Het grootste probleem tijdens de wedstrijden zijn de spreekkoren. [ws-u-e-a-0000000218.p.7.s.3]
'The main problem during the matches are the choirs.'

The predicate nominal in (77) contains both an inherently collective noun and an overt uniqueness marker, and the one in (78) is both topicalized and provided with an overt uniqueness marker. The treebank contains 40 instances which fulfill one or more of the three triggering conditions.

Notice that these are sufficient but not necessary conditions. (40), for instance, repeated in (79), clearly has a collective interpretation, but it does not show any of the tell-tale signs: It is not headed by an inherently collective noun, it does not have an overt uniqueness marker and it is not topicalized.

(79) Politieke tegenstellingen zijn een wezenskenmerk van elke democratie.
'Political contrasts are a defining characteristic of every democracy.'

This suggests that the assignment of the interpretation is also steered by non-linguistic factors. This will be confirmed in sections 5.2.3 and 5.3.

5.2 Triggers of a distributive interpretation

In combinations with a distributive interpretation the predicate nominal is morpho-syntactically singular, but has an underspecified INDEX|NUMBER value. Some examples were already discussed in section 4.2, see (42–43). The purpose of this section is to identify a number of factors which trigger a distributive interpretation.

5.2.1 Quantified subjects

Plural subjects which are introduced by a quantifying determiner have a plural index and, hence, trigger a distributive interpretation. Besides the examples in (42–43), they include those in (80–81).

(80) Niet alle commissarissen zijn werkelijk commissaris. [wiki-7064.p.45.s.5]
'Not all inspectors are really inspector.'

(81) Beide steden waren afwisselend de hoofdstad van het hertogdom. [wiki-154.p.8.s.4]
'Both towns were alternatingly the capital of the duchy.'

The predicate nominals in these combinations tend to be bare singulars, and if they are not, as in (81), the determiner may be omitted without any effect on the grammaticality of the sentence. This is not possible if the nominal has a unique referent.

The distributive interpretation is also triggered by subjects which are introduced by a numeral, as in (82–83).

(82) Vijf Vlamingen op duizend zijn drager van het virus dat hepatitis B veroorzaakt. [wr-p-p-i-0000000011.p.1.s.1]
'Five out of 1000 Flemish people are bearer of the virus that causes hepatitis B.'

(83) Volgens sommige bronnen werden minstens 156 mensen hiervan het slachtoffer. [wr-p-e-i-0000004745.p.5.s.36]
'According to certain sources at least 156 people became the victim of this.'

(82), for instance, does not mean that five out of 1000 Flemish people collectively bear a certain virus, but rather that each one of those five is a bearer of the virus.

A common property of the predicate nominals in (80–83) is that they are preceded and outscoped by the subject. If they are topicalized and, hence, followed by the subject, the combination gets a collective interpretation. It should be added, though, that topicalization of the predicate nominals yields a rather awkward result in (80–83), and that the two observed instances of topicalized PREDCs in sentences with a quantified subject, i.e. (73) and (74), differ from those in this paragraph in the sense that their subject also contains a definite determiner.

5.2.2 Other quantifiers

The quantifying element which triggers the distributive interpretation need not be part of the subject. It can also be a floating quantifier, as in (84–85).

(84) De Arabische staten die onder Brits bewind hadden gestaan werden veelal een monarchie. [wr-p-e-i-0000015007.p.1.s.175]
'The Arab states which had been under British rule, mostly became monarchies.'

(85) We zijn allemaal het slachtoffer van de platonische manier van denken in tweedelingen. [dpc-vla-001161-nl-sen.p.87.s.2]
'We are all victim to the platonic way of thinking in dichotomies.'

(84) does not mean that the Arab states which were part of the British empire now collectively constitute a monarchy, but rather that a number of such states have become monarchies.[23]

The quantifying element may also be a frequency adjunct, such as *vaak* 'often'.

(86) ... dat vrouwen vaak het eerste slachtoffer waren van de politieke instabiliteit en het aanhoudende geweld. [dpc-cam-001017-nl-sen.p.4.s.2]
'... that women were often the first victim of political instability and incessant violence.

(86) is not about a single act of violence which collectively concerns women, but rather about a multitude of such acts, which concern different (sets of) women.

The quantifying element may even be in another clause, as in (87), where the subject in the subordinate clause is an anaphoric pronoun which is bound by a quantified NP in the matrix clause.

(87) Zo zullen [steeds minder jongemannen]$_i$ zichzelf in een volgende generatie ervan kunnen overtuigen dat ze$_i$ "een goede moslim" zijn als ze onschuldige medemensen afmaken. [dpc-ind-001636-nl-sen.p.19.s.4]
'Always fewer youngsters will be able to convince themselves that they are "a good muslim" if they kill innocent people.'

(87) does not mean that fewer youngsters see themselves as collectively constituting a good moslim if they kill innocent people, but rather that fewer youngsters see themselves as good moslim individuals in those conditions.

[23] If the quantifier *veelal* 'mostly' is dropped, the most plausible interpretation is the collective one, in which the various Arab states are understood to become one monarchy.

5.2.3 No overt quantifier, but distributive nonetheless

Sentences can have a distributive interpretation, also if they do not contain any overt quantifier. Some relevant examples are given in (88–89).

(88) Dat betekent niet dat [de initiatiefnemers]$_i$ nu ineens managers zijn. Ze$_i$ zijn en blijven vooral boer. [ws-u-e-a-0000000217.p.26.s.2]
'That does not mean that the initiators are now all of a sudden managers. They are and remain in the first place farmers.'

(89) Overigens zullen de drempels niet gelden voor werknemers uit [Malta en Cyprus]$_i$. [Die eilanden]$_i$ worden per 1 mei óók EU-lidstaat. [ws-u-e-a-0000000043.p.9.s.6]
'The tresholds will not apply to workers from Malta and Cyprus. These islands become EU member states as well on May 1st.'

It is the wider context that triggers the assignment of a distributive interpretation here. The distributive nature of the second sentence in (88), for instance, is favoured by its parallellism with the first sentence, where the predicate nominal is the plural *managers*. In (89) it is a nonlinguistic fact that triggers the assignment of the distributive interpretation: It is because Malta and Cyprus are separate states that *die eilanden* 'those islands' is interpreted as standing for two islands which both become a EU member state. If Malta and Cyprus were two islands which jointly constitute one state, then the collective interpretation would prevail.

5.2.4 Summing up

Distributive interpretations are typically triggered by plural subjects with a quantifying determiner or numeral, but floating quantifiers and frequency adjuncts may have the same effect, and even in the absence of any explicitly quantifying element the most plausible interpretation may be the distributive one. In the treebank we identified 26 instances of unmistakably distributive interpretations.

5.3 In the grey zone

Since the clear-cut cases of collective interpretations and distributive interpretations jointly amount to 66 instances, they cover less than half of the 174 combinations of a plural subject, a plural verb and a singular predicate nominal. Most

of the rest have a collective interpretation, but there are also some which are genuinely ambiguous between a collective and a distributive interpretation. Some examples are given in (90–91).

(90) Verder werden de Romeinse kolonies in Trier en Keulen afzetgebied voor de producten van de Noord-Gallische landbouw en de inheemse ambachten. [wiki-208.p.10.s.1]
'The Roman colonies in Trier and Cologne became a market for the products of the North-Gallic agriculture and the local crafts.'

(91) En opnieuw waren Iraki's die samenwerken met de coalitietroepen het doelwit. [ws-u-e-a-0000000041.p.24.s.2]
'Once again the Iraqi who collabrate with the coalition troops became the target.'

(90) is about two Roman colonies and can either mean that they collectively became a client of Gallic products or that each of them separately became a client. Similarly, (91) is about Iraqis who collaborate with the coalition troops and can either mean that they collectively became a target or that each single one of them became a target.

6 Conclusion

The research on which this article reports has a dual aim. The theoretical aim is to enhance our understanding of the number agreement phenomenon in copular constructions. Building on work in Head-driven Phrase Structure Grammar, and especially on the distinction which it makes between morpho-syntactic agreement and index agreement, we have developed a unified account of the phenomenon, which is made explicit in terms of the general constraint on index agreement in (60) and a small number of independently needed lexical rules, such as (61) and (63). The resulting model copes with a larger range of data than earlier proposals, such as that of Kathol (1999). Moreover, it is straightforwardly extensible to the predicate adjectives of the Romance languages, if one assumes that those adjectives have an index whose NUMBER value is systematically identical to their morpho-syntactic NUMBER value. The methodological aim is to demonstrate how treebanks can be used to guide the formulation of relevant generalizations. For that purpose we rely on tools and resources that have recently become available in the framework of the STEVIN and CLARIN programmes.

References

Augustinus, L., Vandeghinste, V. and Van Eynde, F.(2012), Example-based treebank querying, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, pp. 3161–3167.

Fillmore, C., Lee-Goldman, R. and Rhomieux, R.(2012), The framenet construction, *in* H. Boas and I. Sag (eds), *Sign-based Construction Grammar*, CSLI Publications, Stanford University, pp. 309–372.

Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I. and van der Wouden, T.(2003), CGN syntactische annotatie, Utrecht/Leuven.

Kathol, A.(1999), Agreement and the syntax-morphology interface in HPSG, *in* R. Levine and G. Greene (eds), *Studies in Contemporary Phrase Structure Grammar*, Cambridge University Press, Cambridge, pp. 223–274.

Pollard, C. and Sag, I.(1994), *Head-driven Phrase Structure Grammar*, CSLI Publications and University of Chicago Press, Stanford/Chicago.

Sag, I., Wasow, T. and Bender, E.(2003), *Syntactic theory. A formal introduction. Second Edition*, CSLI Publications, Stanford.

Sauerland, U. and Elbourne, P.(2002), Total reconstruction, PF movement and derivational order, *Linguistic Inquiry* **33**, 283–319.

Spyns, P. and Odijk, J. (eds)(2013), *Essential Speech and Language Technology for Dutch*, Springer, Berlin.

Van Eynde, F.(2003), Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands, Leuven.

Van Eynde, F.(2006), NP-internal agreement and the structure of the noun phrase, *Journal of Linguistics* **42**, 139–186.

Van Eynde, F.(2009), On the copula: from a Fregean to a Montagovian treatment, *in* S. Müller (ed.), *Proceedings of the 16th International Conference on Head-driven Phrase Structure Grammar*, CSLI Publications, Stanford, pp. 359–375.

Van Eynde, F.(2012), On the agreement between predicative complements and their target, *in* S. Müller (ed.), *Proceedings of the 19th International Conference on Head-driven Phrase Structure Grammar*, CSLI Publications, Stanford University, pp. 348–366.

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., Tjong Kim Sang, E. and Vandeghinste, V.(2013), Large scale syntactic annotation of written Dutch: Lassy, *in* P. Spyns and J. Odijk (eds), *Essential Speech and Language Technology for Dutch*, Springer, Berlin, pp. 147–164.

Wechsler, S. and Zlatić, L.(2003), *The many faces of agreement*, CSLI Publications, Stanford.